Analyzing Hate Speech Data along Racial, Gender and Intersectional Axes

Anonymous ACL submission

Abstract

Warning: This work contains strong and offensive language, sometimes uncensored.

To tackle the rising phenomenon of hate speech, efforts have been made towards data curation and analysis. When it comes to analysis of bias, previous work has focused predominantly on race. In our work, we further investigate bias in hate speech datasets along racial, gender and intersectional axes. We identify strong bias against AAE, male and AAE+Male tweets, which are annotated as disproportionately more hateful and offensive than from other demographics. We provide evidence that BERT-based models propagate this bias and show that balancing the training data for these protected attributes can lead to fairer models with regards to gender, but not race.

1 Introduction

Hate Speech. To tackle the phenomenon of online hate speech, efforts have been made to curate datasets (Davidson et al., 2017; Guest et al., 2021; Sap et al., 2020). Since datasets in this domain are dealing with sensitive topics, it is of utmost importance that biases are kept to a (realistic) minimum and that data is thoroughly analyzed before use (Davidson et al., 2019a; Madukwe et al., 2020). In our work, we are contributing to this analysis by uncovering biases along the racial, gender and intersectional axes.

Racial, Gender and Intersectionality Biases. In data collection projects, biases can be introduced in a dataset due to—among other reasons—lack of annotator training or divergence between annotators and user demographics. For example, oftentimes the majority of annotators are white or male (Sap et al., 2020; Founta et al., 2018). An annotator not in the ‘in-group’ may hold (un)conscious biases based on misconceptions about ‘in-group’ speech which may affect their perception of speech from certain communities (O’Dea et al., 2015), leading to incorrect annotations when it comes to dialects the annotators are not familiar with. A salient example of this is annotators conflating African American English (AAE) with offensive or hateful language (Sap et al., 2019).

Intersectionality (Crenshaw, 1989) is a framework for examining how different forms of inequality (for example, racial or gender inequalities) intersect with and reinforce each other. These new social dynamics need to be analyzed both separately and as a whole in order to address challenges faced by the examined communities. For example, a black woman does not face inequality based only on race or only on gender: she faces inequality because of both these characteristics, separately and in conjunction. In this work, we are analyzing not only the racial or gender inequalities in hate speech datasets, but their intersectionality as well.

With research in the area of hate speech, the NLP community aims at protecting target groups and fostering a safer online environment. In this sensitive area, it is pivotal that datasets and models are analyzed extensively to ensure the biases we
are protecting affected communities from do not appear in the data itself, causing further marginalization (for example, by removing AAE speech disproportionately more often).

**Contributions.** In summary, we (i) investigate racial, gender and intersectional bias in three hate speech datasets, Founta et al. (2018); Davidson et al. (2017); Mathew et al. (2021), (ii) examine classifier predictions on existing, general-purpose AAE/SAE and gendered tweets, (iii) identify model bias against AAE, male and AAE+Male (labeled as both AAE and male) tweets, (iv) show that balancing training data for gender leads to fairer models.

## 2 Related Work

Hate speech research has focused on dataset creation (Davidson et al., 2017; Founta et al., 2018; Sap et al., 2020; Guest et al., 2021; Hede et al., 2021; Grimminger and Klinger, 2021) and dataset analysis (Madukwe et al., 2020; Wiegand et al., 2019; Swamy et al., 2019). In our work, we further analyze datasets to uncover latent biases.

It has been shown that data reflects social bias inherent in annotator pools (Waseem, 2016; Al Kuwatly et al., 2020; Davidson et al., 2019a,b). Work has been conducted to identify bias against AAE (Sap et al., 2019; Zhou et al., 2021; Xia et al., 2020) and gender (Excell and Al Moubayed, 2021).

Kim et al. (2020) investigated whether bias along the intersectional axis exists in Founta et al. (2018). While Kim et al. (2020) focused on bias within a single dataset, in our work we generalize to multiple hate speech datasets. We also examine classifier behavior and methods to mitigate this bias.

Research from a sociolinguistic perspective has shown that genders exhibit differences in online text (Gefen and Ridings, 2005) as well as general speech (Penelope Eckbert, 2013). In Bamman et al. (2014) and Bergsma and Van Durme (2013), gender classifiers for English tweets were developed with accuracy of 88% and 85% respectively. In our work, we develop a gender classifier of tweets as well, focusing on precision over recall, leading to a smaller but more accurate sample of gendered data.

## 3 Datasets

Five English datasets were used: three hate speech datasets (DAVIDSON, FOUNTA and HATEXPLAIN), one dataset of tweets labeled for race (GROEWOLD) and one for gender (VOLKOVA).

<table>
<thead>
<tr>
<th></th>
<th>Neutral</th>
<th>Offensive</th>
<th>Hateful</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAVIDSON</strong></td>
<td>0.95</td>
<td>0.95</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>FOUNTA</strong></td>
<td>0.86</td>
<td>0.88</td>
<td>0.37</td>
</tr>
<tr>
<td><strong>HATEXPLAIN</strong></td>
<td>0.69</td>
<td>0.50</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Table 1: F1-score of BERT for each label, evaluated on DAVIDSON, FOUNTA and HATEXPLAIN.

**DAVIDSON.** In Davidson et al. (2017), a hate speech dataset of tweets was collected, labeled for neutral, offensive and hateful language. Hateful language is defined as speech that contains expressions of hatred towards a group or individual on the basis of protected attributes like ethnicity, gender, race and sexual orientation.

**FOUNTA.** In Founta et al. (2018) a crowdsourced dataset of tweets was presented, labeled for normal, abusive and hateful language. To unify definitions, we rename normal to neutral language and abusive to offensive language.

**HATEXPLAIN.** Mathew et al. (2021) presented a dataset from Twitter and Gab1 passages. It has been labeled for normal (neutral), offensive and hateful language.

**GROEWOLD.** In Groenwold et al. (2020) a dataset of African American English and Standard American English tweets was introduced. The AAE tweets come from (Blodgett et al., 2016) and the SAE are direct translations of those tweets provided by annotators.

**VOLKOVA.** Volkova et al. (2013) presented a dataset of 800k English tweets from users with self-identified gender (female/male).

## 4 Experimental Setup

**AAE Classifier.** To classify tweets as AAE or SAE, we used the Blodgett et al. (2016) classifier. Since the number of tweets in our examined datasets was not sufficiently large, we could not use the recommended 0.8 threshold since it did not yield enough results for a confident analysis. We instead fell back to the 0.5 threshold, which can be interpreted as a straightforward classifier of AAE/SAE (whichever class has the highest score is returned).

**Gender Classifier.** To classify tweets as male or female, we finetuned BERT-base2 on Volkova et al. (2013), which includes gender information as self-reported from the tweet authors. We split

---

1 Gab is a social platform that has been known to host far-right groups and rhetoric.
2 https://huggingface.co/bert-base-cased
the dataset into train/dev (50K/25K/25K) and employed a confidence score of 0.8 as the threshold for assigning gender to a tweet. For the tweets with a confidence over the given threshold, precision was 78.4% when classifying tweets as ‘male’ and 79.5% when classifying tweets as ‘female’.

**Hate Speech Classifiers.** For each of the three hate speech datasets (DAVIDSON, FOUNTA, HATEXPLAIN) we finetuned BERT-base. Performance is shown in Table 1. In DAVIDSON and FOUNTA, BERT performs well for neutral and offensive examples, but performs poorly for hateful content. In HATEXPLAIN, BERT overall performs worse, with slightly better performance for neutral and hateful examples over offensive ones.

**Intersectionality.** For our analysis, we classified tweets from all datasets for gender and race.

## 5 Intersectionality Statistics

In Table 2, we present statistics for gender, race and their intersection as found in the three examined hate speech datasets as well as in GROENWOLD and VOLKOVA. We show that no dataset is balanced between AAE and SAE. In FOUNTA and HATEXPLAIN, AAE tweets make up approximately 1/10th of the data. In DAVIDSON, we see stronger representation of AAE, with the AAE tweets being almost twice as many as the SAE tweets. DAVIDSON is also balanced for gender. The other hate speech datasets, while still not balanced, are more balanced for gender than they are for race. FOUNTA has twice as many male than female tweets and HATEXPLAIN has four times as many.

In Table 3, we present a breakdown of protected attributes per class (neutral/offensive/hateful) for DAVIDSON, FOUNTA and HATEXPLAIN. A main takeaway for DAVIDSON and FOUNTA is the imbalance of AAE versus SAE. In SAE, the neutral class makes up 52% of the data for DAVIDSON and 81% for FOUNTA, while the respective numbers for AAE are 3% for DAVIDSON and 13% for FOUNTA.

In HATEXPLAIN, AAE and SAE are more balanced, but there is instead imbalance between genders. For male and female speech, passages are neutral at rates of 43% and 61% respectively. In DAVIDSON, SAE+Female speech is viewed as more offensive than SAE+Male (48% vs. 19%), while in HATEXPLAIN, SAE+Male is more hateful than SAE+Female (34% vs. 16%). Finally, when comparing genders in AAE speech, we see that while AAE+Female contains a larger percentage of offensive tweets (for example, in FOUNTA, 69% vs. 54% and in HATEXPLAIN, 50% vs. 21%), AAE+Male contains disproportionately more hateful speech (in DAVIDSON, 7% vs. 5%, in FOUNTA, 28% vs. 9% and in HATEXPLAIN, 19% vs. 6%).

Overall, AAE and male speech is annotated as more offensive and hateful than SAE and female speech. Further analyzing AAE, AAE+Male is viewed as more hateful than AAE+Female.

### 6 Bias in BERT

We investigate to what extent data bias is learned by BERT. We compare our findings against a dataset balanced for race and gender, to examine whether balanced data leads to fairer models. Namely, we compare a randomly sampled with a balanced set the DAVIDSON dataset. In the balanced set we sample the same number of AAE and SAE tweets (3000) and the same number of male and female tweets (1750). We also include 8000 neutral tweets without race or gender labels. For the randomly sampled set, for a fair comparison, we sampled the same number of tweets as the balanced set. All sampling was stratified to preserve the original label distributions. Results are shown in Table 4.

In the randomly sampled set, there is an imbalance both for gender and race. For gender, while male tweets are more hateful (3% vs. 1%), female tweets are more offensive (71% vs. 63%). For race, AAE is marked almost entirely as offensive (94%), while SAE is split in neutral and offensive (53%)

---

1. Forounta and HateXplain were not considered for this study as they do not contain enough AAE examples to make confident inferences.
2. Experiments were conducted with the entirety of the original dataset with similar results. They are omitted for brevity.

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>SAE</th>
<th>AAE</th>
<th>SAE+Male</th>
<th>AAE+Male</th>
<th>SAE+Female</th>
<th>AAE+Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAVIDSON</td>
<td>2716</td>
<td>2338</td>
<td>3534</td>
<td>8099</td>
<td>1279</td>
<td>1240</td>
<td>3157</td>
<td>1172</td>
</tr>
<tr>
<td>FOUNTA</td>
<td>26307</td>
<td>13615</td>
<td>43330</td>
<td>4177</td>
<td>13486</td>
<td>13257</td>
<td>971</td>
<td>787</td>
</tr>
<tr>
<td>HATEXPLAIN</td>
<td>4509</td>
<td>1103</td>
<td>10368</td>
<td>1103</td>
<td>4145</td>
<td>2376</td>
<td>250</td>
<td>240</td>
</tr>
<tr>
<td>GROENWOLD</td>
<td>587</td>
<td>601</td>
<td>1980</td>
<td>0</td>
<td>587</td>
<td>601</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>VOLKOVA</td>
<td>41164</td>
<td>58836</td>
<td>37874</td>
<td>3755</td>
<td>16243</td>
<td>21031</td>
<td>1843</td>
<td>1912</td>
</tr>
</tbody>
</table>

Table 2: Protected attribute statistics for DAVIDSON, FOUNTA, HATEXPLAIN, GROENWOLD and VOLKOVA.
and 41%). In the SAE subset of tweets, there is an imbalance between genders, with SAE+Female being marked disproportionately more often as offensive than SAE+Male (54% vs. 19%).

### 6.1 Balanced Training

In Table 4, before balancing, 34% of male and 28% of female tweets are marked as neutral. After balancing, these rates are both at 25%. There is an improvement in the intersection of AAE and gender, with the distributions of AAE+Male and AAE+Female tweets converging. For SAE, SAE+Male and SAE+Female distributions converge too, although still far apart. Overall, balanced data improves fairness for gender but not for race, which potentially stems from bias in annotation.

### 6.2 Interpretability with LIME

In Table 5, we show the top contributing words for offensive and hateful predictions in DAVIDSON, FOUNTA and HATEXPLAIN. We see that for AAE, terms such as ‘n****z’ and ‘n***a’ contribute in classifying text as non-neutral even though the terms are part of African American vernacular (Rahman, 2012), showing that this dialect is more likely to be flagged. In non-AAE speech (which includes—but is not exclusive to—SAE), we see the n-word variant with the ‘-er’ spelling appearing more often in various forms, which is correctly picked up by the model as an offensive and hateful term. On both sets, we also see other slurs, such as ‘f*ggots’, ‘moron’ and ‘wetback’ (a slur against foreigners residing in the United States, especially Mexicans) being picked up, showing the model does recognize certain slurs and offensive terms.

### 7 Conclusion

In our work, we analyze racial, gender and intersectional bias in hate speech datasets. We show that tweets from AAE and AAE+Male users are labeled disproportionately more often as offensive. We further show that BERT learns this bias, flagging AAE speech as significantly more offensive than SAE. We perform interpretability analysis using LIME, showing that the inability of BERT to differentiate between variations of the n-word across dialects is a contributing factor to biased predictions. Finally, we investigate whether training on a dataset balanced for race and gender mitigates bias. This method shows mixed results, with gender bias being mitigated more than racial bias. With our work we want to motivate further investigation in model bias not only for the usual gender and racial attributes, but also for their intersection.

<table>
<thead>
<tr>
<th>Table 3: Distribution of protected attribute annotations for neutral/offensive/hateful (N/O/H) examples.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Davidson</td>
</tr>
<tr>
<td>Founta</td>
</tr>
<tr>
<td>HATEXPLAIN</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4: Distribution of predictions for protected attributes on random and balanced datasets based on DAVIDSON. The balanced set is balanced on race (equal number of AAE and SAE tweets) and gender (equal number of female and male tweets). Shown are percentages for neutral/offensive/hateful (N/O/H) predictions.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Random</td>
</tr>
<tr>
<td>Balanced</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5: Top 10 most contributing words for DAVIDSON, FOUNTA and HATEXPLAIN as computed with LIME for hateful predictions.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAVIDSON</strong></td>
</tr>
<tr>
<td>negerize, subhuman, bastards, border, pigfucking, feminist, wetbacks, savages, wetback, jumpers</td>
</tr>
<tr>
<td><strong>FOUNTA</strong></td>
</tr>
<tr>
<td>moron, insult, muslims, aggression, puritan, haters, arabs, coloured, ousted, pedophiles</td>
</tr>
<tr>
<td><strong>HATEXPLAIN</strong></td>
</tr>
<tr>
<td>towelhead, muzzrat, muscum, negociases, niggerette, niglets, musloid, niggerish, niggery, gorilla</td>
</tr>
</tbody>
</table>

In our work, we analyze racial, gender and intersectional bias in hate speech datasets. We show that tweets from AAE and AAE+Male users are labeled disproportionately more often as offensive. We further show that BERT learns this bias, flagging AAE speech as significantly more offensive than SAE. We perform interpretability analysis using LIME, showing that the inability of BERT to differentiate between variations of the n-word across dialects is a contributing factor to biased predictions. Finally, we investigate whether training on a dataset balanced for race and gender mitigates bias. This method shows mixed results, with gender bias being mitigated more than racial bias. With our work we want to motivate further investigation in model bias not only for the usual gender and racial attributes, but also for their intersection.
8 Ethical Considerations

In our work we are dealing with data that can catalyze harm against marginalized groups. We do not advocate for the propagation or adoption of this hateful rhetoric. With our work we wish to motivate further analysis and documentation of sensitive data that is to be used for the training of models (for example, using templates from Mitchell et al. (2019); Bender and Friedman (2018)).

Further, while classifying protected attributes such as race or gender is important in analyzing and identifying bias, care should be taken for the race and gender classifiers to not be misused or abused, in order to protect the identity of users, especially those from marginalized demographics who are more vulnerable to hateful attacks and further marginalization. In our work we only predict these protected attributes for investigative purposes and do not motivate the direct application of such classifiers.

Finally, in our work we only focused on English and a specific set of attributes. Namely, we considered race (African American) and gender. This is a non-exhaustive list of biases and more work needs to be done for greater coverage of languages and attributes.

References


